AUTO-ENCODING EXPLANATORY EXAMPLES

Anonymous authors
Paper under double-blind review

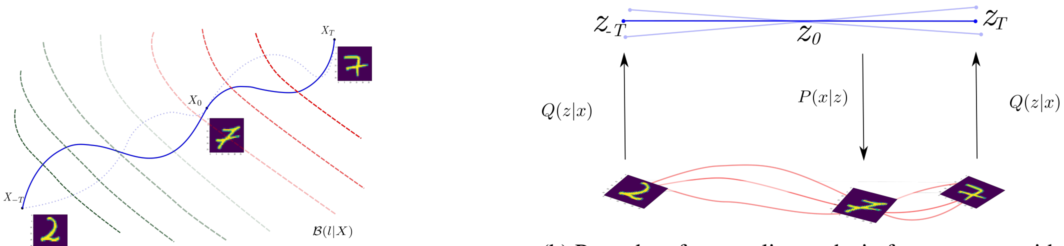
ABSTRACT

In this paper we ask for the main factors that determine a classifier's decision making process and uncover such factors by studying latent codes produced by auto-encoding frameworks. To deliver an explanation of a classifier's behaviour, we propose a method that provides series of examples highlighting semantic differences between the classifier's decisions. These examples are generated through interpolations in latent space. We introduce and formalize the notion of a semantic stochastic path, as a suitable stochastic process defined in feature (data) space via latent code interpolations. We then introduce the concept of semantic Lagrangians as a way to incorporate the desired classifier's behaviour and find that the solution of the associated variational problem allows for highlighting differences in the classifier decision. Very importantly, within our framework the classifier is used as a black-box, and only its evaluation is required.

1 INTRODUCTION

A considerable drawback of the deep classification paradigm is its inability to provide explanations as to why a particular model arrives at a decision. This black-box nature of deep systems is one of the main reasons why practitioners often hesitate to incorporate deep learning solutions in application areas, where legal or regulatory requirements demand decision-making processes to be transparent. A state-of-the-art approach to explain misclassification is saliency maps, which can reveal the sensitivity of a classifier to its inputs. Recent work (Adebayo et al., 2018), however, indicates that such methods can be misleading since their results are at times independent of the model, and therefore do not provide explanations for its decisions. The failure to correctly provide explanations by some of these methods lies in their sensibility to feature space changes, i.e. saliency maps do not leverage higher semantic representations of the data. This motivates us to provide explanations that exploit the semantic content of the data and its relationship with the classifier. Thus we are concerned with the question: *can one find semantic differences which characterize a classifier's decision?*

In this work we propose a formalism that differs from saliency maps. Instead of characterizing particular data points, we aim at generating a set of examples which highlight differences in the decision of a black-box model. Let us consider the task of image classification and assume a misclassification has taken place. Imagine, for example, that a female individual was mistakenly classified as male, or a smiling face was classified as not smiling. Our main idea is to articulate explanations for such misclassifications through sets of semantically-connected examples which link the misclassified image with a correctly classified one. In other words, starting with the misclassified point, we change its features in a suitable way until we arrive at the correctly classified image. Tracking the black-box output probability while changing these features can help articulate the reasons why the misclassification happened in the first place. Now, how does one generate such a set of semantically-connected examples? Here we propose a solution based on a variational auto-encoder framework. We use interpolations in latent space to generate a set of examples in feature space connecting the misclassified and the correctly classified points. We then condition the resulting feature-space paths on the black-box classifier's decisions via a user-defined functional. Optimizing the latter over the space of paths allows us to find paths which highlight classification differences, e.g. paths along which the classifier's decision changes only once and as fast as possible. A basic outline of our approach is given in Fig. 1. In what follows we introduce and formalize the notion of stochastic semantic paths — stochastic processes on feature (data) space created by decoding latent code interpolations. We formulate the corresponding path integral formalism which allows for a Lagrangian formulation of the problem, viz. how to condition stochastic semantic paths on the output



(a) Paths in feature space with black-box classifier level sets.

(b) Procedure for sampling paths in feature space with auto-encoders: interpolations in latent space and decoding of images.

Figure 1: Auto-Encoding Examples Setup: Given a misclassified point x_0 and representatives x_{-T}, x_T, we construct suitable interpolations (stochastic processes) by means of an Auto-Encoder. Sampling points along the interpolations produces a set of examples highlighting the classifier's decision making.

probabilities of black-box models, and introduce an example Lagrangian which tracks the classifier's decision along the paths. We show the explanatory power of our approach on the MNIST and CelebA datasets.

2 Explanations

We are concerned with the problem of explaining a particular decision of a black-box model. Many recent works discuss the roll and provide definitions of explanations in the machine learning context (Doshi-Velez et al., 2017; Gilpin et al., 2018; Abdul et al., 2018; Mittelstadt et al., 2019). Here we follow Ribeiro et al. (2016) and, in broad terms, to explain we mean *to provide textual or visual artifacts that provide qualitative understanding of the relationship between the data points and the model prediction*. Attempts to clarify such a broad notion of explanation require the answers to questions such as (1) *what were the main factors in a decision?*, as well as (2) *would changing a certain factor have changed the decision?* (Doshi-Velez et al., 2017). To provide an answer to such questions, one must be able to define a clear notion of *factors*. One can think of factors as the minimal set of coordinates that allows us to describe the data points. This definition mirrors the behavior and purpose of the variational auto-encoder (VAE) code — by training an auto-encoder one can find a code which describes a particular data point. Our role here is to provide a connection between these latent codes and the classifier's decision. Changes on the code should change the classification decision in a user-defined way. Defining such a code will allow us to formalize the framework required to provide an answer to question (1) and (2) above. Following Ribeiro et al. (2016) we require explanations to be *model-agnostic*, i.e independent of the classifier's inner workings, *interpretable*, and expressing *local fidelity*.

3 Semantics and Example Generation: VAE

Following the discussion above, we use the variational auto-encoder (VAE) formalism (Kingma & Welling, 2013) to introduce a notion of semantics useful to qualitatively explain the decisions of a black-box classifier.

Let us denote the feature (data) space by $\mathcal{X}$ and the latent linear space of codes (describing the data) by $\mathcal{Z}$, where usually $\dim(\mathcal{Z}) \ll \dim(\mathcal{X})$. We consider a latent variable generative model whose distribution $P_\theta(X)$ on $\mathcal{X}$ is defined implicitly through a two-step generation process: one first samples a code Z from a fixed prior distribution $P(Z)$ on $\mathcal{Z}$ and then (stochastically) maps Z to feature space through a (decoder) distribution $P_\theta(X|Z)$, the latter being parametrized by neural networks with parameters θ. This class of models are generically train by minimizing specific distances between the empirical data distribution $P_D(X)$ and the model distribution $P_\theta(X)$. VAE approaches this problem by introducing an encoder distribution $Q_\phi(Z|X)$, parametrized by neural networks with parameters ϕ, which approximates the true posterior distribution $P_\theta(Z|X)$ and minimizing a variational upper bound on the Kullback-Leibler divergence D_{KL} between $P_\theta(X)$ and $P_D(X)$. This bound reads

$$L_{\text{VAE}} = \mathbb{E}_{P_D(X)} \left\{ -\mathbb{E}_{Q_\phi(Z|X)} \left[\log p_\theta(x|z) \right] + D_{\text{KL}} \left(Q_\phi(Z|X), P(Z) \right) \right\}, \tag{1}$$

where $p_\theta(x|z)$ denotes the decoder's density and yields the likelihood function of the data given the code[1]. Once the model is trained one can think of the inferred latent code as containing some high-level description of the input data. Below we will use such inferred code to modify in a controlled way the features of a given input data point.

4 Explaining Through Examples: A Plaintiff Scenario

We define a defendant black-box model $b(l, x)$ as a classifier which yields the probability that the data point $x \in \mathcal{X}$ in feature (data) space belongs to the class $l \in \mathcal{L}$, where $\mathcal{L}$ is a set of classes. Assume the model $b(l, x)$ *is expected to perform* by its users or clients, in a dataset $\mathcal{D} = \{(l_i, x_i)\}$, where $x_i \in \mathcal{X}$ and $l_i \in \mathcal{L}$ is the label that x_i belongs to. [2] Suppose now that the following litigation case emerges. The black-box model b has assigned the data point x_0 to the class l_0. Accordingly, a plaintiff presents a complaint as the point x_0 should have been classified as l_t. Furthermore, assume we are given two additional representative data points x_{-T}, x_T which have been correctly classified by the black-box model to the classes l_{-T}, l_T, respectively — as expected by e.g. the plaintiff, the defendant (if agreed), or the institution upon which the complain or litigation case is presented (say, the court). With this set-up in mind, we propose that an explanation why x_0 was misclassified can be articulated through an *example set* $\mathcal{E} = \{x_{-T}, \ldots, x_0, \ldots, x_T\}$, where $x_t \sim P_\theta(X|Z = z_t)$. Here $P_\theta(X|Z = z_t)$ is a given decoder distribution and the index t runs over *semantic changes* (properly defined below) that highlight classification decisions. This example set constitutes the context revealing how factor changes impact the classification decision (see Section 2). One expects that human oriented explanations are semantic in character. One can understand the expression *bigger eyes will change the classification.* As opposed to changes in some specific pixels [3]. The index t would run over these changes e.g. would make the eyes bigger.

5 Stochastic Semantic Processes and Corresponding Paths

In this section we first formalize the notion of semantic change by introducing the concept of (stochastic) semantic interpolations in feature space $\mathcal{X}$. This will allow us to generate examples which provide *local fidelity*, as the examples are smooth modifications of the latent code associated to the plaintiff data point x_0. We then define a collection of probability measures over semantic paths in $\mathcal{X}$. These measures will be used later in Section 6 to constrain the paths to be explanatory with respect to the classifier's decision.

5.1 Semantic Interpolations

One of the main motivations behind the VAE formalism is the ability of the inferred latent code z to provide semantic high-level information over the data set. If one is to generate examples which have characteristics common to two different data points, say x_0 and x_T from the litigation case, one can perform interpolations between the latent codes of these points, that is z_0 and z_T, and then decode the points along the interpolation. A main observation is that these interpolations in latent space can be used to **induce certain interpolating stochastic processes on feature space**[4] $\mathcal{X}$. We refer to these as *stochastic semantic processes*. In what follows, we first focus on *linear* latent interpolations, i.e.

$$z(t) := t\, z_0 + (1 - t) z_T, \tag{2}$$

and construct an interpolating stochastic semantic process X_t on $\mathcal{X}$ by using the decoder distribution $P_\theta(X|Z = z(t))$. In practice, the generation process of such stochastic interpolations consists then

[1] where the average over $Q_\phi(Z|X)$ is performed using the reparametrization trick (Kingma & Welling, 2013).

[2] Notice that, as a true black-box classifier, one does not know the nature of the true dataset where the model was trained.

[3] the behavior of say adversarial examples

[4] Moreover, under appropriate assumptions on the auto-encoder mappings (P_θ, Q_ϕ) the proposed induced stochastic processes could posses additional properties (e.g. trajectory regularity, controlled moments, etc).

of three steps: (i) sample $Q_\phi(Z|X)$ at the end points x_0 and x_T using the reparametrization trick (Kingma & Welling, 2013), (ii) choose a set of points z_t along the line connecting z_0 and z_T and (iii) decode the z_t by sampling $P_\theta(X|Z = z_t)$. A formal description of this procedure is given below, in subsection 5.2, and an impression of the stochastic process thus constructed is presented in Fig. 1b.

5.2 An Approach via Explicit Family of Measures

We observe that for every sequence of points $\{t_i\}_{i=0}^n$ there is a natural measure on piecewise linear paths starting at $x_0 \in \mathcal{X}$ and terminating at $x_T \in \mathcal{X}$. More precisely, we define the probability of a piecewise linear path $x(t)$ with nodes $x_1, x_2 \ldots, x_n \in \mathcal{X}$ as

$$dP_{t_0,\ldots,t_n}(x(t)) := \int_{\mathcal{Z}} \int_{\mathcal{Z}} \left(\prod_{i=1}^{n} p_\theta(x_i|z(t_i)) \right) q_\phi(z_0|x_0) q_\phi(z_T|x_T)\, dz_0\, dz_T, \tag{3}$$

where q_ϕ, p_θ label the densities of Q_ϕ, P_θ, respectively, and where $z(t)$ is defined by eq. (2) [5].

In other words, for every pair of points x_0 and x_T in feature space, and its corresponding code samples $z_0 \sim Q_\phi(Z|X = x_0)$ and $z_T \sim Q_\phi(Z|X = x_T)$, the decoder $P_\theta(X|Z)$ induces a measure over the space of paths $\{x(t)|x(0) = x_0, x(T) = x_T\}$. Formally speaking, the collection of measures $d\mathcal{P}_{t_0,\ldots,t_n}$ given by different choices of points $\{t_i\}_{i=0}^n$ in (3) defines a family of consistent measures (cf. Definition 2 in the Appendix, Subsection D.1). This implies that these different measures are assembled into a stochastic process on feature space $\mathcal{X}$ over the continuous interval $[0, T]$:

Proposition 1. *The collection of measures prescribed by (3) induces a corresponding continuous-time stochastic process. Moreover, under appropriate reconstruction assumptions on the auto-encoder mappings P_θ, Q_ϕ, the sample paths are interpolations, that is, start and terminate respectively at x_0, x_T almost surely.*

The statement goes along the lines of classical results on existence of product measures. For the sake of completeness we provide all the necessary technical details in the Appendix, Subsection D. Another important remark is that the stochastic semantic process construction in Proposition 1 is just one way to define such a process — there are other natural options, e.g. in terms of explicit transition kernels or Itô processes.

6 Principle of Least Semantic Action

Having described a procedure to sample stochastic semantic processes in $\mathcal{X}$, we need to discover auto-encoding mappings (P_θ, Q_ϕ) that give rise to reasonable and interesting stochastic paths. Specifically, to generate examples which are able to explain the defendant black-box model $b(l, x)$ in the current litigation case (Section 4), one needs to ensure that semantic paths between the data points x_0 and x_T *highlight classification differences*, i.e. classifications of the model along this path are far apart in the plaintiff pair of labels. Thus, to design auto-encoding mappings P_θ, Q_ϕ accordingly, we propose an optimization problem of the form

$$\min_{\theta,\phi} S_{P_\theta,Q_\phi}[X_t], \tag{4}$$

where X_t is a stochastic semantic process and S_{P_θ,Q_ϕ} is an appropriately selected functional that extracts certain features of the black-box model $b(l, x)$.

The minimization problem (4) can be seen in the context of Lagrangian mechanics. For a given stochastic semantic process X_t, and given initial and final feature “states" x_0 and x_T, we introduce the following function, named the *model-b semantic Lagrangian*

$$\mathcal{L} : [0, 1] \times \mathcal{X} \times \mathcal{X} \to \mathbb{R}, \quad (t, x_0, x_T) \mapsto \mathcal{L}[X_t, x_0, x_T], \tag{5}$$

which gives rise to the *semantic model action*:

$$S[X_t] := \int_0^T \mathcal{L}[X_t, x_0, x_T]dt. \tag{6}$$

[5] We remark that the integral (eq. 3) is, moreover, finite, if, for example, the densities p_θ are bounded with respect to z.

In mechanics, the optimization given by suitable Lagrangians delivers physically meaningful paths, e.g. those specified by the equations of motion (Landau & Lifshitz, 2013). In our case, a guiding intuition is that the semantic Lagrangian should reflect how the black-box model takes decisions along the path[6] X_t, starting at x_0 and ending at x_T. In this way, the minimization of the semantic action (i.e. finding minimizing paths X_t) should make such classification aspects prominent along the example set.

Our problem, viz. to find encoding mappings P_θ, Q_ϕ which yield explainable semantic paths with respect to a black-box model, is then a constrain optimization problem whose total objective function we write as

$$L(\theta, \phi) := L_{\text{VAE}}(\theta, \phi) + \lambda\, \mathbb{E}_{dP[x(t)]} S[x(t)], \tag{7}$$

where L_{VAE} is given by eq. (1), $S[x(t)]$ corresponds to the Lagrangian action and λ is an hyper parameter controlling the action' scale. The average over the paths (Majumdar, 2007; Feynman & Hibbs, 1965) is taken with respect to the stochastic paths and the corresponding measure $dP[x(t)]$ from Proposition 1, that is, the path integral

$$\mathbb{E}_{dP[x(t)]} S[(x(t))] = \int \mathcal{L}[x(t), x_0, x_T] dP[x(t)] \approx \frac{1}{nK} \sum_{k}^{K} \sum_{t}^{n} \mathcal{L}[x_t^k, x_0, x_T], \tag{8}$$

where x_t^k labels the tth point along the the kth path, sampled as described in Section 5, n is the number of points on each path, K is the total number of paths, and the estimator on the right hand side corresponds to an explicit average over paths[7].

Algorithm 1: PATH Auto-Encoder

```
Data: Dataset $\mathcal{D} = \{(x_i, l_i)\}$ Litigation case
      $(x_{-T}, x_T, l_0, x_0, l_t, \mathcal{B}(l|x))$
Encoder $P_\theta(x|z)$, Decoder $Q_\phi(z|x)$
while $\phi$ and $\theta$ not converged do
    Draw $\{x_1, ..., x_n\}$ from the training set
    Calculate Auto-Encoder Loss $L_{\text{VAE}}(\theta, \phi)$
    Sample Litigation Codes
    $z_{-T} \sim Q_\phi(Z|x_{-T})$, $z_0 \sim Q_\phi(Z|x_0)$,
      $z_T \sim Q_\phi(Z|x_T)$
    Generate Latent Interpolations
    $t_j^k \sim$ Sort(Uniform(0,1))
    $z_j^k = z_{-T} \times t_j^k + z_0 \times (1 - t_j^k)$
    Sample $k$ Paths in Feature Space
    $x_t^k \sim P_\theta(X|z_j^k)$
    Evaluate Semantic Action for each path $k$
    and average over $k$
    $L_{\mathcal{S}} = \mathbb{E}_{d\mathcal{P}[x(t)]}[\mathcal{S}(x(t))]$
    Update $P_\theta$ and $Q_\phi$ by descending:
      $L_{\text{VAE}}(\theta, \phi) + L_{\mathcal{S}}(\theta, \phi)$
end
return $P_\theta$, $Q_\phi$
```

In practice, both L_{VAE} and the action term are optimized simultaneously. Note that the VAE loss function L_{VAE} is trained on the entire data set on which the black-box performs. The action term, in contrast, only sees the x_0 and x_T points. This can be seen explicitly in Algorithm 1, which shows an overview of the auto-encoder pair training algorithm. Let us finally note that, drawing analogies with the adversarial formalism (Goodfellow et al., 2015), the defendant black-box model plays the role of a fixed discriminator, not guiding the example generation, but the interpolations among these examples.

6.1 The Choice of Lagrangians

There are plenty of options for Lagrangian functionals that provide reasonable (stochastic) example-paths — roughly speaking, we attempt to define an objective value for a certain subjective notion of explanations. In what follows we illustrate one particular such Lagrangian[8]

Minimum Hesitant Path

We want to find an example path such that the classifier's decisions along it changes as quickly as possible, as to highly certain regions in $\mathcal{X}$. In

[6] For example, a valid Lagrangian should reflect whether the the decisions along the path were taken with sufficient certainty, or whether the probability of the decisions were gradually changing.

[7] Note that, as mention in Sec. 5, one must resort to the reparametrization trick to sample from Q_ϕ and efficiently evaluate the gradients of the action term. Note also that Proposition 1 tells us that different choices of the discrete (approximation) grids in the t integration are qualitatively related to the same underlying stochastic interpolation process.

[8] For further suggestions concerning the Lagrangian choice we refer to the Appendix, Subsection D.3. There we also compute the corresponding Euler-Lagrange equations that shed further light on the solutions of the variational problem — moreover, their solution could be used in the training process as well.

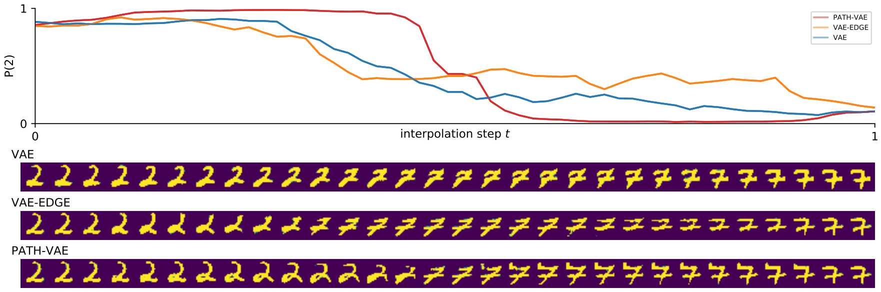


Figure 2: Probability Paths for the litigation case $l_0 = 2, l_T = 7$. Y axis corresponds to classification probability and x axis corresponds to interpolation index. Interpolation images for a specific paths are presented below the x axis.

other words, the path is forced to stay in regions where the black-box produces decisions with maximum/minimum probability. An intuitive way to enforce this is via the simple Lagrangian

$$\mathcal{L}_1(x(t), x_0, x_T) := -\left(b(l_T, x(t)) - b(l_0, x(t))\right)^2, \tag{9}$$

where l_0, l_T are the labels of the litigation case in question. Roughly speaking, given the appropriate initial conditions, the paths that minimize the action associated to $\mathcal{L}_1$ are paths that attempt to keep $\mathcal{L}_1$ close to 1 over almost the entire interpolation interval.

OTHER REGULARIZERS

Additionally we require $b(l_T, x(t))$ to be a monotonous function along the interpolating path $x(t)$. Furthermore, in accordance with Proposition 1 we require certain level of reconstruction at the end points. To enforce these conditions we introduce the regularizers r_m, r_e which are described in detail in subsection D.4 of the Appendix.

The total objective function is therefore

$$L(\theta, \phi) := L_{\text{VAE}}(\theta, \phi) + \lambda\, \mathbb{E}_{dP[x(t)]} S_1[x(t)] + \lambda_m r_m + \lambda_e r_e, \tag{10}$$

where $\lambda, \lambda_m, \lambda_e$ are hyper-parameters and S_1 is the action associated to the minimum hesitant Lagrangian $\mathcal{L}_1$ in eq. (9).

7 EXPERIMENTAL RESULTS

We evaluate our method in two real-world data sets: MNIST, consisting of 70k Handwriting digits, (LeCun, 1998) and CelebA (Liu et al., 2015), with roughly 203k images. We use a vanilla version of the VAE (Kingma & Welling, 2013) with Euclidean latent spaces $\mathcal{Z} = \mathbb{R}^{d_z}$ and an isotropic Gaussian as a prior distribution $P(Z) = \mathcal{N}(Z|0, I_{d_z})$. We used Gaussian encoders, i.e. $Q_\phi(Z|X) = \mathcal{N}(Z|\mu_\phi(X), \Sigma_\phi(X))$, where μ_ϕ, σ_ϕ are approximated with neural networks of parameters ϕ, and Bernoulli decoders $P_\theta(X|Z)$. We compare the standard VAE, VAE-EDGE (VAE augmented with the edge loss r_e) and PATH-VAE (our full model, eq. (10)). The black-box classifier $b(l, x)$ is defined as a deep network with convolutional layers and a final soft-max output layer for the labels. Details of the specifics of the architectures as well as training procedure are left to the Appendix.

For MNIST we studied a litigation case wherein $l_{-T}, l_T = 2, 7$ and $l_0 = 2$, whereas its true label (i.e. that of x_0) is $l_t = 7$ (see Section 4). The results are presented in Fig. 2. VAE delivers interpolations which provide uninformative examples, i.e. the changes in the output probability $b(l_0, x)$ *cannot* be associated with changes in feature space. In stark contrast, PATH-VAE causes the output probability to change abruptly. This fact, together with the corresponding generated examples, allows us to propose explanations of the form: what makes the black-box model classify an image in the path as *two* or *seven*, is the shifting up of the lower stroke in the digit *two* as to coincide with the middle bar of the digit *seven*. Similarly, the upper bar of the digit *seven* (especially the upper left part) has a significant decision weight.

In order to provide a more quantitative analysis we demonstrate the capability of our methodology to control the path action while retaining the reconstruction capacity. Hereby, we use not only the VAE as the underlying generative model, but also Wasserstein Auto-Encoder (WAE) (Bousquet et al., 2017) and Adversarial Auto-Encoder (AAE) (Goodfellow et al., 2015), i.e. we simply change L_{VAE} in eq. (7) with the loss of WAE or AAE. The theoretical details and corresponding architectures are presented in the Appendix. We present, in Fig. 3, the action values defined over random litigation end pairs (x_{-T}, x_T). The PATH version of the model indeed yields lower action values. Furthermore, these models tend to reduce the variance within the different paths. This is expected since there is one path that minimizes the action, hence, the distribution will try to arrive at this specific path for all samples. In order to compare with other explanation models, we define a saliency map with the interpolations obtained in our methodology. We defined the *interpolation saliency* as the sum over the differences between interpolation images weighted with the probability change of the black-box classifier through the interpolation path. We see in Fig. 4 the comparisons among different methods. While the standard methods only show local contributions to a classification probability, our saliency maps show the minimum changes that one is to apply to the dubious image in order to change the decision to the desired label. Our approach reveals that the curvature of the lower bar is decisive to be classified as a two, while the style of the upper bar is important to be classified as a seven. Further, we provide a sanity check analysis (Adebayo et al., 2018) by studying the rank correlation between original saliency map and the one obtained for a randomized layers of the black-box classifier, shown in Fig. 5. As desired, our proposed saliency map decorrelates with the randomized version.

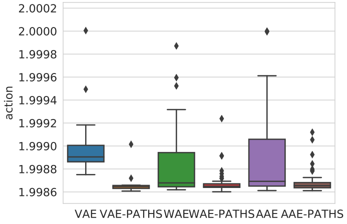


Figure 3: Average action for Minimum Hesitant Lagrangian. PATHS-architectures trained to minimize semantic action.

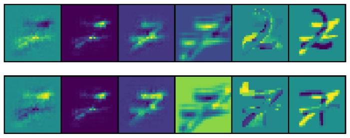


Figure 4: Saliency Maps Comparison: From left to right column: Vanilla Gradients, Smooth Gradients, Guided BackProp, Grad CAMP, Interpolations, Difference with Representative. Upper row corresponds to $l_{-T} = 2$, lower row to $l_T = 7$.

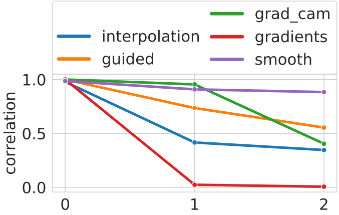


Figure 5: Rank correlation between the original explanation and the randomized explanation derived up to that layer.

For the CelebA dataset we use a black-box classifier based on the ResNet18 architecture (He et al., 2016). We investigate two specific misclassifications. In the first case, a smile was not detected (Fig. 6 a). Here we only interpolate between the misclassified image (left) and a correctly classified one (right), of the same person. Interpolations obtained by VAE are not informative: specific changes in feature space corresponding to changes in the probability cannot be detected since the latter changes rather slowly over the example path. This observation also holds true for the VAE-EDGE model, except that the examples are sharper. Finally, our PATH-VAE model yields a sharp change in the probability along with a change of the visible teeth (compare the third and fifth picture in the example path), revealing that this feature (i.e. teeth visibility) could constitute, from a human standpoint, a decisive factor in the probability of detecting a smile for the given black-box model. It is important to note that these observations represent one of many possible path changes which could change the classifier decision. This is constrained by the current realization and representative end points. The important result is that our methods are able to shape the behavior of the classifier along the path. Further experimental examples are provided in Section C of the Appendix.

8 Related Work

The bulk of the explanation literature for deep/black-box models relies on input dependent methodologies. Gradient Based approaches (Simonyan et al., 2013; Erhan et al., 2009) derive a sensibility

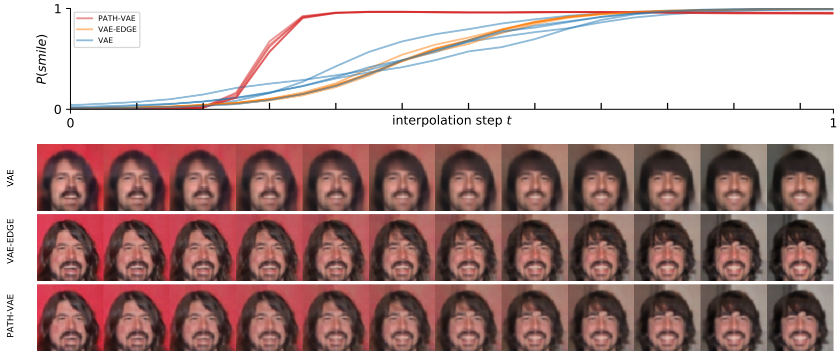


Figure 6: Probability Paths for the case of detecting a smile in images of celebrities. Y axis corresponds to classification probability and x axis corresponds to interpolation index. Interpolation images for a specific paths are presented below the x axis. The images are vertically aligned with a corresponding tick in the x-axis determining the interpolation index of the image

score for a given input example and class label by computing the gradient of the classifier with respect to each input dimension. Generalizations of this approach address gradient saturation by incorporating gradients' values in the saliency map (Shrikumar et al., 2017) or integrating scaled versions of the input (Sundararajan et al., 2017). Ad hoc modifications of the gradient explanation via selection of the required value (Springenberg et al., 2015),(Zeiler & Fergus, 2014), as well as direct studies of final layers of the convolutions units of the classifiers (Selvaraju et al., 2016), are also provided. In contrast to gradient based approaches, other categories of explanatory models rely on *reference based approaches* which modify certain inputs with uninformative reference values (Shrikumar et al., 2017). Bayesian approaches treat inputs as hidden variables and marginalize over the distribution to obtain the saliency of the input (Zintgraf et al., 2017). More recent generalizations exploit a variational Bernoulli distribution over the pixels values (Chang et al., 2018). Other successful methodologies include substitution of black-box model with locally interpretable linear classifiers. This is further extended to select examples from the data points in such a way that the latter reflect the most informative components in the linear explanations, (Ribeiro et al., 2016). Studies of auto-encoder interpolations seek to guarantee reconstruction quality. In (Arvanitidis et al., 2018) the authors characterize latent space distortions compared to the input space through a stochastic Riemannian metric. Other solutions include adversarial cost on the interpolations such as to improve interpolation quality compared to the reconstructions, (Berthelot et al., 2018). Examples which are able to deceive the classifier's decisions have been widely studied in the framework of adversarial examples (Goodfellow et al., 2015). These methodologies, however, do not provide interpretable explanations or highlight any semantic differences that lead to the classifier's decisions. Finally, the Auto-Encoder framework can also naturally be seen as a tool for **dimensionality reduction**. Geometrically speaking, assuming that the data set approximately lies along a manifold embedded in feature space $\mathcal{X}$, one can interpret the encoder, decoder as the coordinate map (chart) and its inverse. From this point of view, our approach above translates to finding coordinate charts with additional constraints on mapping the segments from z_0 to z_T to appropriate (stochastic) curves between x_0 and x_T.

9 Conclusion

In the present work we provide a novel framework to explain black-box classifiers through examples obtained from deep generative models. To summarize, our formalism extends the auto-encoder framework by focusing on the interpolation paths in feature space. We train the auto-encoder, not only by guaranteeing reconstruction quality, but by imposing conditions on its interpolations. These conditions are such that information about the classification decisions of the model $\mathcal{B}$ is encoded in the example paths. Beyond the specific problem of generating explanatory examples, our work formalizes the notion of a stochastic process induced in feature space by latent code interpolations, as well as quantitative characterization of the interpolation through the semantic Lagrangian's and actions. Our methodology is not constrained to a specific Auto-Encoder framework provided that mild regularity conditions are guaranteed for the auto-encoder.

REFERENCES

Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pp. 582:1–582:18, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5620-6. doi: 10.1145/3173574.3174156. URL `http://doi.acm.org/10.1145/3173574.3174156`.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity Checks for Saliency Maps. In *Advances in Neural Information Processing Systems*, pp. 9525–9536, 2018.

Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent Space Oddity: on the Curvature of Deep Generative Models. *International Conference on Learning Representations*, 2018.

Christian Bär and Frank Pfäffle. Wiener Measures on Riemannian Manifolds and the Feynman-Kac Formula. *Preprints des Instituts für Mathematik der Universität Potsdam 1*, 2012.

David Berthelot, Colin Raffel, Aurko Roy, and Ian J. Goodfellow. Understanding and Improving Interpolation in Autoencoders via an Adversarial Regularizer. *CoRR*, abs/1807.07543, 2018.

Olivier Bousquet, Sylvain Gelly, Ilya Tolstikhin, Carl Johann Simon-Gabriel, and Bernhard Schölkopf. From Optimal Transport to Generative Modeling: the VEGAN cookbook. Technical report, 2017.

Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. Explaining Image Classifiers by Counterfactual Generation. *arXiv preprint arXiv:1807.08024 [cs.CV]*, 2018.

Manfredo P. do Carmo. *Differential Geometry of Curves and Surfaces*. Prentice Hall, 1976. ISBN 978-0-13-212589-5.

Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Stuart Schieber, James Waldo, David Weinberger, and Alexandra Wood. Accountability of AI Under the Law: The Role of Explanation. *CoRR*, abs/1711.01134, 2017.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing Higher-Layer Features of a Deep Network. *University of Montreal*, 1341(3):1, 2009.

Richard Feynman and Albert Hibbs. Quantum Mechanics and Path Integrals. *International Series in Pure and Applied Physics, McGraw-Hill*, 1965.

L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 80–89, Oct 2018. doi: 10.1109/DSAA.2018.00018.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. *International Conference on Learning Representations*, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Lev Davidovich Landau and Evgenii Mikhailovich Lifshitz. *Course of Theoretical Physics*. Elsevier, 2013.

Yann LeCun. The MNIST Database of Handwritten Digits. NEC Research Institute, 1998.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Satya N Majumdar. Brownian Functionals in Physics and Computer Science. In *The Legacy Of Albert Einstein: A Collection of Essays in Celebration of the Year of Physics*, pp. 93–129. World Scientific, 2007.

Brent Mittelstadt, Chris Russell, and Sandra Wachter. Explaining explanations in ai. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pp. 279–288, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-6125-5. doi: 10.1145/3287560.3287574. URL `http://doi.acm.org/10.1145/3287560.3287574`.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why Should I Trust You?: Explaining the Predictions of any Classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144. ACM, 2016.

Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-CAM: Why did you say that? *arXiv preprint arXiv:1611.07450*, 2016.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning Important Features Through Propagating Activation Differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3145–3153, 2017.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *CoRR*, abs/1312.6034, 2013.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for Simplicity: The All Convolutional Net. *International Conference on Learning Representations (Workshop)*, 2015.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3319–3328, 2017.

Michael Taylor. Partial Differential Equations II: Qualitative Studies of Linear Equations. *Applied Mathematical Sciences 116, Springer-Verlag New York*, 2011.

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein Auto-Encoders. In *International Conference on Learning Representations*, 2018.

Matthew D Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks. In *European Conference on Computer Vision*, pp. 818–833. Springer, 2014.

Luisa M. Zintgraf, Taco S. Cohen, Tameem Adel, and Max Welling. Visualizing Deep Neural Network Decisions: Prediction Difference Analysis. *5th International Conference on Learning Representations, 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

A Appendix

B Models Training Details

B.1 A Gaussian CNN Encoder and CNN Decoder: MNIST and CelebA

There was no preprocessing on the 28x28 MNIST images. The models were trained with up to 100 epochs with mini-batches of size 32 - we remark that in most cases, however, acceptable convergence occurs much faster, e.g. requiring up to 15 epochs of training. Our choice of optimizer is Adam with learning rate $\alpha = 10^{-3}$. The weight of the KL term of the VAE is $\lambda_{kl} = 1$, the *path* loss weight is $\lambda_p = 10^3$ and the *edge* loss weight is $\lambda_e = 10^{-1}$. We estimate the *path* and *edge* loss during training by sampling 5 paths, each of those has 20 steps.

Encoder Architecture

$$\begin{aligned} x \in \mathbb{R}^{28\times28\times3} &\to \text{Conv}_{64} \to \text{BN} \to \text{ReLU} \\ &\to \text{Conv}_{128} \to \text{BN} \to \text{ReLU} \\ &\to \text{Conv}_{256} \to \text{BN} \to \text{ReLU} \to \text{FC}_{8} \end{aligned}$$

Decoder Architecture

$$\begin{aligned} z \in \mathbb{R}^8 &\rightarrow \text{FC}_{4\times4\times512} \\ &\rightarrow \text{FSConv}_{256} \rightarrow \text{BN} \rightarrow \text{ReLU} \\ &\rightarrow \text{FSConv}_{128} \rightarrow \text{BN} \rightarrow \text{ReLU} \\ &\rightarrow \text{FSConv}_{64} \rightarrow \text{Sigmoid} \end{aligned}$$

Both the encoder and decoder used fully convolutional architectures with 3x3 convolutional filters with stride 2. Conv$_k$ denotes the convolution with k filters, FSConv$_k$ the fractional strides convolution with k filters (the first two of them doubling the resolution, the third one keeping it constant), BN denotes batch normalization, and as above ReLU the rectified linear units, FC$_k$ the fully connected layer to $\mathbb{R}^k$.

The pre-processing of the CelebA images was done by first taking a 140x140 center crop and then resizing the image to 64x64. The models are trained with up to 100 epochs and with mini-batches of size 128. Our choice of optimizer is Adam with learning rate $\alpha = 10^{-3}$. The weight of the KL term of the VAE is $\lambda_{kl} = 0.5$, the *path* loss weight is $\lambda_p = 0.5$ and the *edge* loss weight is $\lambda_e = 10^-3$. We estimate the *path* and *edge* loss during training by sampling 10 paths, each of those has 10 steps.

Encoder Architecture

$$\begin{aligned} x \in \mathbb{R}^{64\times64\times3} &\rightarrow \text{Conv}_{64} \rightarrow \text{BN} \rightarrow \text{ReLU} \\ &\rightarrow \text{Conv}_{128} \rightarrow \text{BN} \rightarrow \text{ReLU} \\ &\rightarrow \text{Conv}_{256} \rightarrow \text{BN} \rightarrow \text{ReLU} \\ &\rightarrow \text{Conv}_{512} \rightarrow \text{BN} \rightarrow \text{ReLU} \rightarrow \text{FC}_{500} \end{aligned}$$

Decoder Architecture

$$\begin{aligned} z \in \mathbb{R}^{500} &\rightarrow \text{FC}_{4\times4\times512} \\ &\rightarrow \text{FSConv}_{256} \rightarrow \text{BN} \rightarrow \text{ReLU} \\ &\rightarrow \text{FSConv}_{128} \rightarrow \text{BN} \rightarrow \text{ReLU} \\ &\rightarrow \text{FSConv}_{64} \rightarrow \text{Sigmoid} \end{aligned}$$

Both the encoder and decoder used fully convolutional architectures with 3x3 convolutional filters with stride 2. Conv$_k$ denotes the convolution with k filters, FSConv$_k$ the fractional strides convolution with k filters (the first two of them doubling the resolution, the third one keeping it constant), BN denotes batch normalization, and as above ReLU the rectified linear units, FC$_k$ the fully connected layer to $\mathbb{R}^k$.

C Further Results

C.1 MNIST

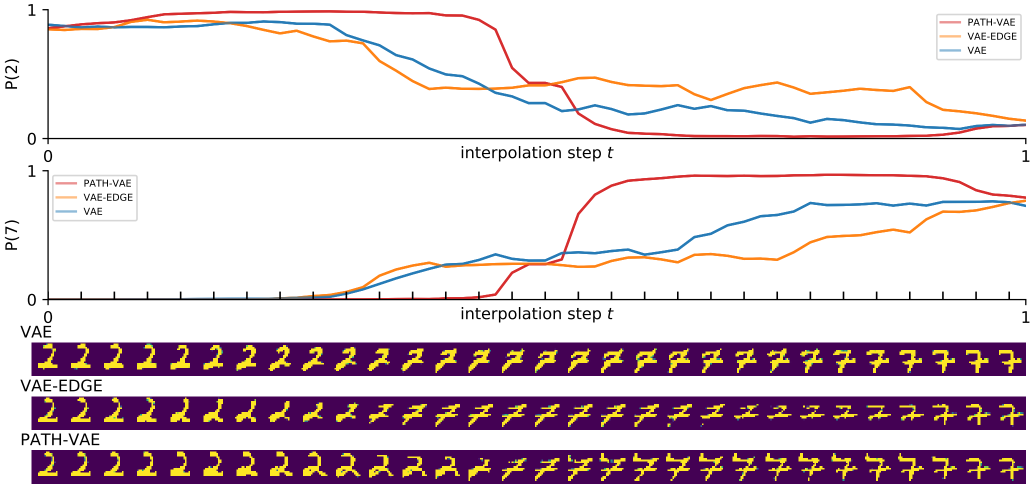


Figure 7: Interpolation between 2 and 7. It is seen that the Path-VAE interpolation optimizes both probabilities (P(2) and P(7)) according to the chosen Lagrangian - in this case the minimum hesitant $\mathcal{L}_1$.

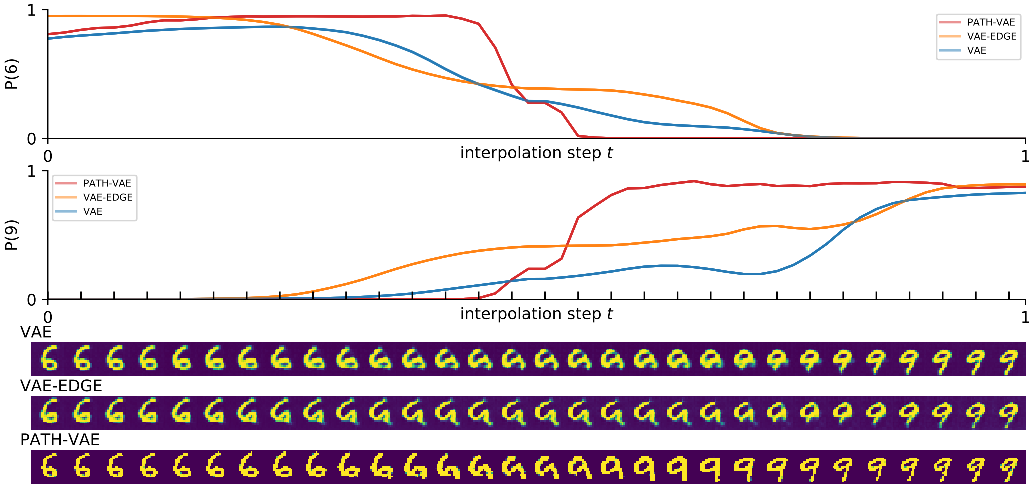


Figure 8: Interpolation between 6 and 9. The Path-VAE interpolation appears to emphasize the "opening and closing" of the loop.

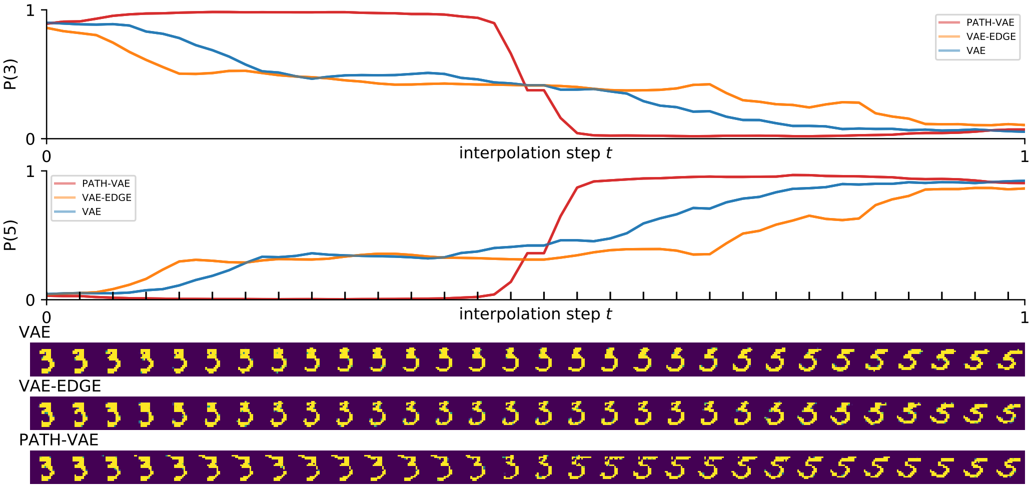


Figure 9: Interpolation between 3 and 5. The translation of the "upper bar" is a prominent feature of the Path-VAE interpolation.

C.2 CELEBA

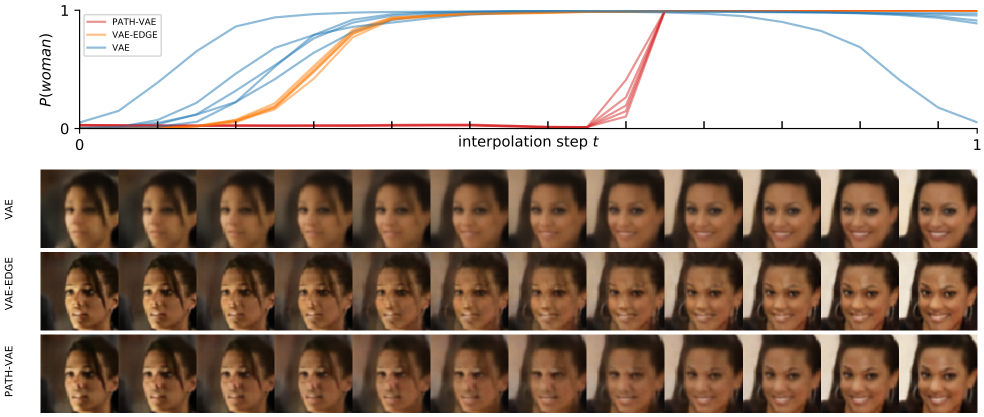


Figure 10: Probability Paths for the case of detecting the sex in images of celebrities. Y axis corresponds to classification probability and x axis corresponds to interpolation index. Interpolation images for a specific paths are presented below the x axis. The images are vertically aligned with a corresponding tick in the x-axis determining the interpolation index of the image

D Stochastic Semantic Processes: Proof of Proposition 1

Briefly put, the construction we utilize makes use of the well-known notion of consistent measures, which are finite-dimensional projections that enjoy certain restriction compatibility; afterwards, we show existence by employing the central extension result of Kolmogorov-Daniell.

D.1 Collections of Consistent Measures

We start with a couple of notational remarks.

Definition 1. *Let S, F be two arbitrary sets. We denote*

$$S^F := \{f : F \to S\}, \tag{11}$$

that is, the set of all maps $F \to S$.

Definition 2. *Let $(S, \mathcal{B})$ be a measurable space and let $G \subseteq F \subseteq [0, T]$ for some positive number T. We define the restriction projections $\pi_{F,G}$ by*

$$\pi_{F,G} : S^F \to S^G, \quad f \in S^F \mapsto f|_G \in S^G. \tag{12}$$

Moreover, for each $F \subseteq [0, T]$ the restriction projections induce the σ-algebra $\mathcal{B}^F$ which is the smallest σ-algebra on S^F so that all projections

$$\pi_{F,\{t\}} : S^F \to S^{\{t\}} \cong S, \quad \forall t \in F, \tag{13}$$

are measurable. In particular, the projections $\pi_{F,G}$ are measurable with respect to $\mathcal{B}^F, \mathcal{B}^G$.

Definition 3. *Let us denote by $\operatorname{Fin}([0, T])$ the set of all finite-element subsets of $[0, T]$. A collection of finite measures $\{(\mu_F, \mathcal{B}^F), F \in \operatorname{Fin}([0, T])\}$ is called consistent if it is push-forward compatible with respect to the restriction projection mappings, i.e.*

$$(\pi_{F,G})_* \, \mu_F = \mu_G, \quad \forall F, G \in \operatorname{Fin}([0, T]), G \subseteq F. \tag{14}$$

Here

$$(\pi_{F,G})_* \, \mu_F(A) := \mu_F(\pi^{-1}_{F,G}(A)), \quad \forall A \in \mathcal{B}^G. \tag{15}$$

Proposition 2. *Let $F = \{0 \leq t_1 < t_2 < \cdots < t_n \leq T\} \in \operatorname{Fin}([0, T])$ be an arbitrary finite set. The mapping*

$$\mu_F(A) := \int \chi_A(x_1, x_2, \ldots, x_n) \left(\prod_{i=1}^{n} p_\theta(x_i|z_i) \right) q_\phi(z_0|x_0) q_\phi(z_T|x_T) dz_0 dz_T dx_1 \ldots dx_n \tag{16}$$

defines a consistent collection of finite measures.

Proof. Let us fix

$$F_1 := \{0 \leq t_1 < t_2 < \cdots < t_n \leq T\} \in \operatorname{Fin}([0, T]), \tag{17}$$

$$F_2 := \{0 \leq t_1 < t_* < t_2 < \cdots < t_n \leq T\} \in \operatorname{Fin}([0, T]), \tag{18}$$

Without loss of generality, it suffices to check consistency for the pair (F_1, F_2). We have

$$(\pi_{F_1,F_2})_* \mu_{F_2}(A) = \mu_{F_2}\left(\pi^{-1}_{F_1,F_2}(A)\right) \tag{19}$$

$$= \int \chi_{\pi^{-1}_{F_1,F_2}(A)}(x_1, s, x_2, \ldots, x_n) \left(\prod_{i=1}^{n} p_\theta(x_i|z_i) \right) \tag{20}$$

$$\times \, p_\theta(s|z_{t_*}) q_\phi(z_0|x_0) q_\phi(z_T|x_T) ds dz_0 dz_T dx_1 dx_2 \ldots dx_n \tag{21}$$

$$= \int \chi_A(x_1, x_2, \ldots, x_n) \left(\prod_{i=1}^{n} p_\theta(x_i|z_i) \right) q_\phi(z_0|x_0) q_\phi(z_T|x_T) dz_0 dz_T dx_1 \ldots dx_n \tag{22}$$

$$= \mu_{F_1}(A), \tag{23}$$

where we have used L^1-finiteness and integrated out the s variable via Fubini's theorem. Note also, that by the definitions above

$$\chi_{\pi^{-1}_{F_1,F_2}(A)}(x_1, s, x_2, \ldots, x_n) = \chi_A(x_1, x_2, \ldots, x_n). \tag{24}$$

for any fixed $s \in \mathcal{X}$. □

We briefly recall the following classical result due to Kolmogorov and Daniell:

Theorem 1 (Theorem 2.11, Bär & Pfäffle (2012))**.** *Let $(S, \mathcal{B}(S))$ be a measurable space with S being compact and metrizable and let I be an index set. Assume that for each $J \in \operatorname{Fin}(I)$ there exists a measure μ^J on $S^J, \mathcal{B}^J$, such that the following compatibility conditions hold:*

$$\mu^{J_1} = \mu^{J_2} \circ \pi_{J_1}^{-1}, \quad \forall J_1 \subseteq J_2 \in \operatorname{Fin}(I). \tag{25}$$

Here $\pi_{J_1} : S^{J_2} \to S^{J_1}$ denotes the canonical projection (obtained by restriction).

Then, there exists a unique measure μ on $(S^I, \mathcal{B}^I)$ such that for all $J \in \operatorname{Fin}(I)$ one has

$$\mu \circ \pi_J^{-1} = \mu^J. \tag{26}$$

We recall that a well-known way to construct the classical Wiener measure and Brownian motion is precisely via the aid of Theorem 1 (Taylor (2011)). We are now in a position to construct the following stochastic process.

Proposition 3. *There exists a continuous-time stochastic process $X_t : [0, T] \to \mathbb{R}^D$ satisfying*

$$\mathbb{P}((X_{t_1}, X_{t_2}, \dots, X_{t_n}) \in A) = \int \chi_A(x_1, x_2, \dots, x_n) \tag{27}$$

$$\times \left(\prod_{i=1}^{n} p_\theta(x_i|z_i) \right) q_\phi(z_0|x_0) q_\phi(z_T|x_T) dx_1 \dots dx_n. \tag{28}$$

$$\tag{29}$$

Moreover, for small positive numbers ϵ, δ we have $X_0 \in B_\delta(x_0)$ with probability at least $(1 - \epsilon)$, provided the reconstruction error of encoding/decoding process is sufficiently small. In particular, if x_0 stays fixed after the application of encoder followed by decoder, then $X_0 = x_0$ almost surely. A similar statement holds also for the terminal point X_t and x_T respectively.

Proof. By applying Theorem 1 to the collection of consistent finite measures prescribed by Proposition 2 we obtain a measure μ on the measurable space $(S^{[0,T]}, \mathcal{B}^{[0,T]})$. Considering the probability space $(S^{[0,T]}, \mathcal{B}^{[0,T]}, \mu)$ we define stochastic process

$$X_t := \pi_{[0,T],\{t\}} : S^{[0,T]} \to S. \tag{30}$$

It follows from the construction and the Theorem of Kolmogorov-Daniell that $\mathbb{P}\left((X_{t_1}, X_{t_2}, \dots, X_{t_n}) \in A\right)$ is expressed in the required way. This shows the first claim of the statement.

Now, considering a small ball $B_\delta(x_0)$ we have

$$\mathbb{P}(X_0 \in B_\delta(x_0)) = \int \chi_{B_\delta(x_0)}(x) p_\theta(x|z_0) q_\phi(z_0|x_0) q_\phi(z_T|x_T) dx dz_0 dz_T \tag{31}$$

$$= \int \chi_{B_\delta(x_0)}(x) p_\theta(x|z_0) q_\phi(z_0|x_0) dx dz_0 \tag{32}$$

$$:= R(x_0, \chi_{B_\delta(x_0)}). \tag{33}$$

Here, the function $R(x^*, U)$ measures the probability that the input x^* is decoded in the set U. Thus, if the reconstruction error gets smaller, R converges to 1. This implies the second statement.

Finally, if we assume that the auto-encoder fixes x_0 in the sense above, we similarly get

$$\mathbb{P}(X_0 = x_0) = \int \chi_{\{x_0\}}(x) p_\theta(x|z_0) q_\phi(z_0|x_0) q_\phi(z_T|x_T) dx dz_0 dz_T \tag{34}$$

$$= \int \chi_{\{x_0\}}(x) p_\theta(x|z_0) q_\phi(z_0|x_0) dx dz_0 \tag{35}$$

$$= \delta_{x_0}(\chi_{\{x_0\}}) \tag{36}$$

$$= 1. \tag{37}$$

□

D.2 Concerning the Regularity of Sample Paths

An important remark related to the the variational problem (4) is the following: one could develop plenty of meaningful functionals $\mathcal{S}_{P_\theta, Q_\phi}$ that involve taking velocities or higher derivatives - thus one is supposed to work over spaces of curves with certain regularity assumptions. However, as stated above we are working over stochastic paths X_t whose regularity is, in general, difficult to guarantee. A straightforward way to alleviate this issue is to consider a "smooth" version of the curve X_t - e.g. by sampling X_t through a decoder with controllable or negligible variance or by means of an appropriate smoothing. Furthermore, one could also approach such stochastic variational analysis via Malliavin calculus - however, we do not pursue this direction in the present work.

We now briefly discuss a few remarks about the regularity of the stochastic semantic process from Proposition 1. First, we state a well-known result of Kolmogorov and Chentsov:

Theorem 2 (Theorem 2.17, Bär & Pfäffle (2012))**.** *Let (M, ρ) be a metric measure space and let $X_t, t \in [0, T]$ be a stochastic process. Suppose that there exists positive numbers a, b, C, ϵ with the property*

$$\mathbb{E}\left[\rho(X_s, X_t)^a\right] \leq C|t-s|^{(1+b)}, \quad \forall s, t, |s-t| < \epsilon \tag{38}$$

Then, there exists a version $Y_t, t \in [0, T]$ of the stochastic process X_t whose paths are α-Hölder continuous for any $\alpha \in (0, b/a)$.

Thus, roughly speaking, an estimate on $\mathbb{E}\left[\rho(X_s, X_t)^a\right]$ can be regarded as a measure of the extent to which Theorem 2 fails. To give an intuitive perspective, let us consider the stochastic process given by Proposition 1 and, considering only the points $X_s, X_{s+\delta}$ for a small positive number δ, let us write the expectation in (38) as:

$$\int\int\int\int \|x_{s+\delta} - x_s\| \, p_\theta(x_{s+\delta}|z_{s+\delta}) p_\theta(x_s|z_s) q_\phi(z_0|x_0) q_\phi(z_T|x_T) dx_s dx_t dz_0 dz_T, \tag{39}$$

where we have used the standard Euclidean distance. To estimate the integral further, let us for simplicity assume that the encoder is deterministic and the decoder is defined via a Gaussian Ansatz of the type $\mu(z) + \sigma(z) \otimes \epsilon$ for a normal Gaussian variable ϵ. Thus the last integral can be written as:

$$\int\int \frac{\|x_{s+\delta} - x_s\|}{(2\pi)^n \sqrt{|\Sigma_{s+\delta}||\Sigma_s|}} \exp\left(-\frac{1}{2}[(x_{s+\delta} - \mu(z_{s+\delta}))^T \Sigma_{s+\delta}^{-1}(x_{s+\delta} - \mu(z_{s+\delta})) \tag{40}$$

$$+(x_s - \mu(z_s))^T \Sigma_s^{-1}(x_s - \mu(z_s))]\right) dx_s dx_t, \tag{41}$$

where we denote the covariance matrix at time s by Σ_s. Now, if $\Sigma_{s+\delta}$ becomes sufficiently small as δ converges to 0, then the exponential factor will dominate and thus (38) holds. In other words, Hölder regularity of the process is verified provided that $p_\theta(x|z)$ becomes localized in x and converges to a Dirac measure (similarly to the case of the heat kernel propagator and Brownian motion). From this point of view, the variance of the decoder can be considered as an indicator of how far the stochastic process is from being Hölder continuous.

Below we discuss two other stochastic process constructions, one of which is built upon Itô diffusion processes and enjoys further path-regularity properties.

D.3 Further Semantic Lagrangians and Associated Euler-Lagrange Equations

We briefly recall that, among other aspects, Lagrangian theory suggests a framework for optimization of functionals (Lagrangians) defined over appropriate function spaces. Critical points of Lagrangians are identified by means of the corresponding Euler-Lagrange equations (Landau & Lifshitz (2013)). To obtain the Euler-Lagrange equations for the Lagrangians in (9, 45, 50) we compute in a straightforward manner the first variation

$$\frac{\delta \mathcal{S}_i}{\delta x}[\phi] := \frac{d}{d\epsilon}|_{\epsilon=0} \mathcal{S}_i(x(t) + \epsilon\phi(t)), \quad i = 1, 2, 3, \tag{42}$$

where $\phi : [0, T] \to T\mathcal{X}$ is a compactly supported deformation [9]. This produces the following conditions:

[9] By $T\mathcal{X}$ we mean the tangent bundle of $\mathcal{X}$.

$$\frac{d}{d\epsilon}|_{\epsilon=0}\mathcal{S}_1(x(t)+\epsilon\phi(t)) = -\frac{d}{d\epsilon}|_{\epsilon=0}\left[\int_0^T \left(\mathcal{B}(l_T|x(t)+\epsilon\phi(t)) - \mathcal{B}(l_0|x(t)+\epsilon\phi(t))\right)^2 dt\right] \quad (43)$$

$$= -2\int_0^T f\langle\nabla f, \phi\rangle dt = -2\int_0^T \sum_{i=1}^{\dim\mathcal{X}} f\,\phi_i\,\partial_i f dt, \quad (44)$$

where we have denoted $f := (\mathcal{B}(l_T|x(t)) - \mathcal{B}(l_0|x(t)))$ and differentiated under the integral sign. The notation $\langle\cdot,\cdot\rangle$ denotes the standard Euclidean scalar product. Requiring that the first variation vanishes for every choice of deformation ϕ implies that either f or ∇f vanishes.

Minimum Decision Path

Contrary to the previous Lagrangian $\mathcal{L}_1$ where the minimizers force almost instantaneous jumps of $\mathcal{B}_\theta$, one might prefer a path that illustrates a uniform change of $\mathcal{B}_\theta$ - such a gradual change might be implemented by requesting that $\mathcal{B}_\theta$ changes linearly along $x(t)$. To this end, we introduce

$$\mathcal{S}_2(x(t), x_0, x_t) := \int_0^T \left(\langle\nabla\mathcal{B}(l_T|x(t)), \dot{x}(t)\rangle - \frac{1}{T}\right)^2 dt := \int_0^T \mathcal{L}_2(x(t), x_0, x_t)dt. \quad (45)$$

Computing the first variation as above, one gets:

$$\frac{d}{d\epsilon}|_{\epsilon=0}\mathcal{S}_2(x(t)+\epsilon\phi(t)) = \frac{d}{d\epsilon}|_{\epsilon=0}\left[\int_0^T \left(\langle\nabla\mathcal{B}(l_T|x(t)+\epsilon\phi(t)), \dot{x}(t)+\epsilon\dot{\phi}(t)\rangle - \frac{1}{T}\right)^2 dt\right] \quad (46)$$

$$= 2\int_0^T \left(\langle\nabla\mathcal{B}(l_T|x(t)), \dot{x}(t)\rangle - \frac{1}{T}\right)\big(\big\langle\nabla^2\mathcal{B}(l_T|x(t)(\phi(t)), \dot{x}(t)\big\rangle \quad (47)$$

$$+\langle\nabla\mathcal{B}(l_T|x(t)), \phi(t)\rangle\big)\,dt, \quad (48)$$

where we have denoted the Hessian by ∇^2 and, as above, the notation $\langle\cdot,\cdot\rangle$ denotes the standard Euclidean scalar product.

Now, assuming that $\mathcal{B}$ is non-degenerate in the sense that the second factor in the integrand is not identically vanishing for all choices of a deformation ϕ (i.e. the Hessian and gradient of $\mathcal{B}$ satisfy the above relations for all t), one sees that the critical points satisfy

$$\frac{d}{dt}\mathcal{B}(l_T|x(t)) = \frac{1}{T}, \quad \forall t\in[0,T]. \quad (49)$$

Such a condition resembles the conservation of angular momentum given, for instance, by the classical formula of Clairaut in the case of surfaces of revolution.

Minimum Transformation Path

Another meaningful Lagrangian construction is given by following the geometry of $\mathcal{B}$ itself and attempting to find paths that are close to being gradient-descent lines. This can be embodied by defining

$$\mathcal{S}_3(x(t), x_0, x_t) := \int_0^T \|\nabla\mathcal{B}(l_T|x(t)) - \alpha\dot{x}(t)\|^2 dt := \int_0^T \mathcal{L}_3(x(t), x_0, x_t), \quad (50)$$

where α is a suitably chosen positive constant describing the extent to which the stochastic path should follow the geometry of $\mathcal{B}$. Computing the variation w.r.t. $x(t)$ one easily sees that

$$\frac{d}{d\epsilon}|_{\epsilon=0}\mathcal{S}_3(x(t)+\epsilon\phi(t)) = \frac{d}{d\epsilon}|_{\epsilon=0}\left[\int_0^T \|\nabla\mathcal{B}(l_T|x(t)+\epsilon\phi(t)) - \alpha\left(\dot{x}(t)+\epsilon\dot{\phi}(t)\right)\|^2 dt\right] \quad (51)$$

$$= 2\int_0^T \left\langle\nabla\mathcal{B}(l_T|x(t)) - \alpha\dot{x}(t), \nabla^2\mathcal{B}(l_T|x(t))(\phi(t)) - \alpha\dot{\phi}(t)\right\rangle dt, \quad (52)$$

Assuming the Hessian is not identically vanishing along the curve, the critical points of the variational problem are given by the condition

$$(\nabla \mathcal{B})\left(l_T | x(t)\right) = \alpha \dot{x}(t). \tag{53}$$

In addition to following the geometry of the black box $\mathcal{B}$, one could also impose a natural condition that the stochastic paths minimize distances on the manifold in feature space that the auto-encoder pair induces. We recall from basic differential geometry that the image of the decoder as a subset of the feature space is a submanifold with a Riemannian metric g induced by the ambient Euclidean metric in the standard way (for background we refer to do Carmo (1976)). In the simple case of a deterministic auto-encoder, one can think of g as the matrix $J^T J$ where J denotes the Jacobian of the decoder - thus g gives rise to scalar product $g(X, Y) := X J^T J Y$. In the stochastic case, one can use suitable approximations to obtain g in a similar manner - e.g. in Arvanitidis et al. (2018) the authors decompose the decoder into a deterministic and a stochastic part, whose Jacobians J_1, J_2 are summed as $J_1^T J_1 + J_2^T J_2$ to obtain the matrix g.

Now, having Riemannian structure (i.e. the notion of a distance) on the data submanifold, geodesic curves naturally arise as minimizers of a suitable distance functional, namely:

$$\mathcal{S}_4(x(t), x_0, x_t) := \int_0^T \|\dot{x}(t)\|_g dt, \tag{54}$$

where the norm $\|\cdot\|_g$ is computed with respect to the Riemannian metric g, that is $\sqrt{g(\cdot,\cdot)}$. We note that the utilization of geodesics for suitable latent space interpolations was thoroughly discussed in Arvanitidis et al. (2018).

D.4 Other Regularizers

As mentioned already, we would like that classifier's probabilities change in a monotonous fashion along the paths - these paths are preferable in the sense that they provide examples following a particular trend along the disputed labels. We enforce such monotonic behaviour along the paths with the term

$$r_m := \frac{1}{K(n-1)} \sum_k^K \sum_i^{n-1} \min(0, b(l_T, x_i^k) - b(l_T, x_{i+1}^k)), \tag{55}$$

with n the number of points along the path and K the number of paths.

Further, and in accordance with Proposition 1, one can also require that the auto-encoder reconstructs the endpoints with sufficiently large accuracy. We enforce this requirement with the edge term $r_e := \sum_i \left(|b(l_i, x_i)) - b(l_i, \tilde{x_i})| + c(x_i, \tilde{x_i})\right), \quad i = 0, T, -T,$where c measures the reconstruction error[10] and $\tilde{x_i} \sim P_\theta(X|Z = z_i)$, with $z_i \sim Q_\phi(Z|X = x_i)$ and x_i the data points at $i = 0, T, -T$.

D.5 Wasserstein

In contrast to VAE, within the WAE framework Tolstikhin et al. (2018) one only needs to be able to sample from $Q_\phi(Z|X)$ and $P_\theta(X|Z)$ — i.e. their density is not needed. WAE is trained by minimizing a (penalized) optimal transport divergence Bousquet et al. (2017) — the Wasserstein distance, between the input data distribution $P_D(X)$ and the implicit latent variable model $P_\theta(X)$. As in VAE, the latter is defined by first sampling Z from $P(Z)$ and then mapping Z to X through the decoder $P_\theta(X|Z)$. The loss function of WAE is given by

$$L_{\text{WAE}} = \mathbb{E}_{P_D(X)} \mathbb{E}_{Q_\phi(Z|X)} \left[c\left(X, P_\theta(X|Z)\right)\right] + \lambda D_Z\left(Q_\phi(Z), P(Z)\right), \tag{56}$$

where c is a distance function and D_Z is an arbitrary divergence between the prior $P(Z)$ and the agregate posterior $Q_\phi(Z) = \mathbb{E}_{P_D(X)}\left[Q_\phi(Z|X)\right]$, weighted by a positive hyperparameter λ. Minimizing Equation 56 corresponds to minimizing the Wasserstein distance if the decoder is deterministic (i.e. $P_\theta(X|Z = z) = \delta_{g_\theta(z)} \forall z \in \mathcal{Z}$, with the map $g_\theta : \mathcal{Z} \to \mathcal{X}$) and the distance term is optimized. If the decoder is stochastic Equation 56 yields an upper bound on the Wasserstein

[10] In our experiments we computed c either by means of the cross-entropy or the Euclidean norm.

distance Bousquet et al. (2017). In the present work, we use two different divergences D_Z.

GAN Based $D_{\mathcal{Z}}$ Here we calculate the Jensen Shannon divergence through an adversarial game, where a discriminator separates true points sampled from the prior P_Z from fake ones sampled from the aggregated posterior. Notice that this corresponds to the classical adversarial cost Goodfellow et al. (2015) performed in the latent space.
MMD-based $D_{\mathcal{Z}}$ For a postive-definite reproducing kernel $k : \mathcal{Z} \times \mathcal{Z} \to \mathcal{R}$ we use the maximum mean discrepancy MMD:

$$MMD_k(P_Z, Q_Z) = || \int_{\mathcal{Z}} k(z, \cdot) dP_Z(z) - \int_{\mathcal{Z}} k(z, \cdot) dQ_Z(z) ||_{\mathcal{H}_{\mathcal{Z}}} \tag{57}$$

In our experiments we choose a squared cost function $c(x, y) = ||x - y||_2^2$ and refer to the Wasserstein with JS divergence for $D_{\mathcal{Z}}$ as Adversarial Auto Encoders (AAEs) [11] whereas the model trained with the mean discrepancy divergence is denoted in the main text as WAE. We use the inverse multi quadratic kernel $Kkx, y) = \frac{C}{C+||x-y||_2^2}$

[11] the equivalence is stablished in Tolstikhin et al. (2018)